5-9-2013

The GenoChip: A New Tool for Genetic Anthropology

Eran Elhaik

Elliott Greenspan

Sean Staats

Thomas Krahn

Chris Tyler-Smith

See next page for additional authors

Recommended Citation

Elhaik, E., Greenspan, E., Staats, S., Krahn, T., Tyler-Smith, C., Xue, Y., Tofanelli, S., Cucca, F., Pagani, L., Jin, L., Li, H., Schurr, T. G., Greenspan, B., Wells, R., & Genographic Consortium (2013). The GenoChip: A New Tool for Genetic Anthropology. *Genome Biology and Evolution, 5* (5), 1021-1031. https://doi.org/10.1093/gbe/evt066

The GenoChip: A New Tool for Genetic Anthropology

Abstract
The Genographic Project is an international effort aimed at charting human migratory history. The project is nonprofit and non-medical, and, through its Legacy Fund, supports locally led efforts to preserve indigenous and traditional cultures. Although the first phase of the project was focused on uniparentally inherited markers on the Y-chromosome and mitochondrial DNA (mtDNA), the current phase focuses on markers from across the entire genome to obtain a more complete understanding of human genetic variation. Although many commercial arrays exist for genome-wide single-nucleotide polymorphism (SNP) genotyping, they were designed for medical genetic studies and contain medically related markers that are inappropriate for global population genetic studies. GenoChip, the Genographic Project's new genotyping array, was designed to resolve these issues and enable higher resolution research into outstanding questions in genetic anthropology. The GenoChip includes ancestry informative markers obtained for over 450 human populations, an ancient human (Saqqaq), and two archaic hominins (Neanderthal and Denisovan) and was designed to identify all known Y-chromosome and mtDNA haplogroups. The chip was carefully vetted to avoid inclusion of medically relevant markers. To demonstrate its capabilities, we compared the F_{ST} distributions of GenoChip SNPs to those of two commercial arrays. Although all arrays yielded similarly shaped (inverse J) F_{ST} distributions, the GenoChip autosomal and X-chromosomal distributions had the highest mean F_{ST}, attesting to its ability to discern subpopulations. The chip performances are illustrated in a principal component analysis for 14 worldwide populations. In summary, the GenoChip is a dedicated genotyping platform for genetic anthropology. With an unprecedented number of approximately 12,000 Y-chromosomal and approximately 3,300 mtDNA SNPs and over 130,000 autosomal and X-chromosomal SNPs without any known health, medical, or phenotypic relevance, the GenoChip is a useful tool for genetic anthropology and population genetics.

Keywords
genetic anthropology, GenoChip, Genographic Project, population genetics, AimsFinder, haplogroups

Disciplines
Anthropology | Biological and Physical Anthropology | Genetic Structures | Social and Behavioral Sciences

Author(s)
Eran Elhaik, Elliott Greenspan, Sean Staats, Thomas Krahn, Chris Tyler-Smith, Yali Xue, Sergio Tofanelli, Francesco Cucca, Luca Pagani, Li Jin, Hui Li, Theodore G. Schurr, Bennett Greenspan, R. Spencer Wells, and Genographic Consortium

GBE

The GenoChip: A New Tool for Genetic Anthropology

Eran Elhaik[1], Elliott Greenspan[2], Sean Staats[2], Thomas Krahn[2], Chris Tyler-Smith[3], Yali Xue[3], Sergio Tofanelli[4], Paolo Francalacci[5], Francesco Cucca[6], Luca Pagani[3,7], Li Jin[8], Hui Li[8], Theodore G. Schurr[9], Bennett Greenspan[2], R. Spencer Wells[10,*], and the Genographic Consortium

[1]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health

[2]Family Tree DNA, Houston, Texas

[3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

[4]Department of Biology, University of Pisa, Italy

[5]Department of Natural and Environmental Science, Evolutionary Genetics Lab, University of Sassari, Italy

[6]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy

[7]Division of Biological Anthropology, University of Cambridge, United Kingdom

[8]MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, People's Republic of China

[9]University of Pennsylvania

[10]National Geographic Society, Washington, DC

*Corresponding author: E-mail: spwells@ngs.org.

Accepted: April 24, 2013

Abstract

The Genographic Project is an international effort aimed at charting human migratory history. The project is nonprofit and nonmedical, and, through its Legacy Fund, supports locally led efforts to preserve indigenous and traditional cultures. Although the first phase of the project was focused on uniparentally inherited markers on the Y-chromosome and mitochondrial DNA (mtDNA), the current phase focuses on markers from across the entire genome to obtain a more complete understanding of human genetic variation. Although many commercial arrays exist for genome-wide single-nucleotide polymorphism (SNP) genotyping, they were designed for medical genetic studies and contain medically related markers that are inappropriate for global population genetic studies. GenoChip, the Genographic Project's new genotyping array, was designed to resolve these issues and enable higher resolution research into outstanding questions in genetic anthropology. The GenoChip includes ancestry informative markers obtained for over 450 human populations, an ancient human (Saqqaq), and two archaic hominins (Neanderthal and Denisovan) and was designed to identify all known Y-chromosome and mtDNA haplogroups. The chip was carefully vetted to avoid inclusion of medically relevant markers. To demonstrate its capabilities, we compared the F_{ST} distributions of GenoChip SNPs to those of two commercial arrays. Although all arrays yielded similarly shaped (inverse J) F_{ST} distributions, the GenoChip autosomal and X-chromosomal distributions had the highest mean F_{ST}, attesting to its ability to discern subpopulations. The chip performances are illustrated in a principal component analysis for 14 worldwide populations. In summary, the GenoChip is a dedicated genotyping platform for genetic anthropology. With an unprecedented number of approximately 12,000 Y-chromosomal and approximately 3,300 mtDNA SNPs and over 130,000 autosomal and X-chromosomal SNPs without any known health, medical, or phenotypic relevance, the GenoChip is a useful tool for genetic anthropology and population genetics.

Key words: genetic anthropology, GenoChip, Genographic Project, population genetics, AimsFinder, haplogroups.

Introduction

Apportionment of human genetic variation has long established that all living humans are related via recent common ancestors who lived in sub-Saharan Africa some 200,000 years ago (Cann et al. 1987). The world outside Africa was settled over the past 50,000–100,000 years (Henn et al. 2010) when the descendents of our African forebears spread out to populate other continents (Cavalli-Sforza 2007).

This "Out-of-Africa" hypothesis, backed by archeological findings (Klein 2008) and genetic evidence (Stringer and Andrews 1988; Laval et al. 2010), describes a major dispersal of anatomically modern humans that completely replaced local archaic populations outside Africa, although a scenario involving Europeans and West Africans admixing with extinct hominins was also proposed (Plagnol and Wall 2006). Remarkably, recent studies proposed evidence for two such archaic admixture (interbreeding) events, one with Neanderthals in Europe and eastern Asia (Green et al. 2010) and the second with Denisovans in Southeast Asia and Oceania (Reich et al. 2011), though the extent of the hybridization remains questionable (Eriksson and Manica 2012). Overall, the recurrent migrations, admixture, and interbreeding events shaped the autosomes of modern populations into mosaics of ancient and recent alleles harbored in haplotypes that vary in size but not in the building blocks themselves. These subtle differences in autosomal allele frequency between populations together with uniparental markers provide genetic data with the potential to obtain evidence of mixing and migration of human populations.

The advent of microarray single-nucleotide polymorphism (SNP) technology that revolutionized human population genetics and broadened our understanding of genetic diversity largely skipped genetic anthropology for three main reasons: first, only a handful of the estimated 5,000–6,000 indigenous population groups (Burger and Strong 1990; Fardon 2012) were genotyped and studied, which may limit the phylogeographic resolution of the findings. Second, the plethora of genetic markers obtained from different genotyping platforms has resurrected the "empty matrix" problem, whereby populations from different studies can barely be compared due to the low overlap of these platforms. Finally, genotyping costs remained prohibitively high and unjustified for genetic anthropology, as the commercial genotyping platforms, by large, do not accommodate ancestry informative markers (AIMs). Furthermore, these arrays are enriched in trait- or disease-related markers, which prompt a host of psychological, social, legal, political, and ethical concerns from the individual to the population and global levels (Royal et al. 2010).

The first phase of The Genographic Project focused on reconstructing human migratory paths through the analysis of uniparentally inherited markers on the Y-chromosome and mitochondrial DNA (mtDNA). The success of the project in both inferring details of human migratory history (e.g., Balanovsky et al. 2011; Schurr et al. 2012) and attracting over half a million public participants interested in tracing their genetic ancestry has prompted entrepreneurs to offer multiple self-test kits that provide information ranging from disease risk and life-style choices (e.g., diet) to genetic ancestry (Wolinsky 2006). Some of these solutions have been criticized for making deceptive health-related claims and providing limited and imprecise answers regarding ancestry (Royal et al. 2010). The concerns about ancestry reporting were not unjustified, as these entrepreneurs adopted the commercial genotyping platforms that were fraught with medically informative markers, depleted of AIMs, and overall yielded biased measures of genetic diversity (Albrechtsen et al. 2010).

Although uniparental arrays do not suffer from the aforementioned predicaments, they are limited in that they represent only a smaller and more ancient portion of our history and ignore our remaining ancestors whose contribution to our genome was more recent and substantial. In contrast, assessment of the spatial and temporal patterns of genetic variation in the rest of the genome coupled with data obtained from other disciplines can provide more information of our ancestors. However, autosomal-driven studies attempting to discern markers informative to genetic anthropology from those having medical relevance often met with legal or ethical obstacles and failed to attract participants who remained concerned about the sharing and potential exploitation of their medical information (Royal et al. 2010). These constraints render all commercial genotyping arrays unsuitable for genetic anthropology, including the Human Origins array (Lu et al. 2011) that contains coding and medically related markers.

To facilitate high-quality research in genetic anthropology without obtaining health, trait, or medical information, we resolved to develop a novel genotyping array—which we call the GenoChip. Our goals were to 1) design a state of the art SNP array dedicated solely to genetic anthropology, 2) validate its accuracy, 3) evaluate its abilities to discern populations compared with alternative arrays, and 4) demonstrate its performances on worldwide populations.

Materials and Methods

Genotype Data Retrieval

AIMs were obtained from 15 studies (Yang et al. 2005; Price et al. 2007, 2008; Halder et al. 2008; Tian et al. 2008, 2009; Florez et al. 2009; Kosoy et al. 2009; McEvoy et al. 2009, 2010; Nassir et al. 2009; Henn et al. 2011; Kidd et al. 2011).

Genotype data for thousands of samples from over 300 worldwide populations were obtained from 15 public and private collections (Conrad et al. 2006; Reich et al. 2009; Silva-Zolezzi et al. 2009; Teo et al. 2009; Xing et al. 2009, 2010; Altshuler et al. 2010; Behar et al. 2010; Hunter-Zinck et al. 2010; Rasmussen et al. 2010, 2011; Chaubey et al. 2011; Hatin et al. 2011; Henn et al. 2011; Yunusbayev et al. 2012) and the FamilyTreeDNA collection. To study gene flow from apes, ancient hominins, and modern humans, we used the data set of 257,000 high-quality autosomal SNPs assembled by Reich et al. (2010).

SNP Validation

To cross-validate the GenoChip's autosomal genotypes, we genotyped 168 samples from 14 worldwide populations of the 1000 Genomes Project including Americans of African

ancestry from Southwest United States (ASW), Americans of Mexican ancestry from Los Angeles, CA (MEX), Utah residents with Northern and Western European ancestry from UT (CEU), England and Scotland British (GBR), Finnish from Finland (FIN), Gujarati Indians from Houston, TX (GIH), Han Chinese from Beijing, China (CHB), Iberians from Spain (IBS), Italians from Tuscany, Italy (TSI), Japanese from Tokyo, Japan (JPT), Kinh from Ho Chi Minh City, Vietnam (KHV), Luhya in Webuye, Kenya (LWK), Peruvians from Lima, Peru (PEL), and Yoruba in Ibadan, Nigeria (YRI). The concordance rate between GenoChip and the 1000 Genomes Project genotypes was calculated as the proportion of genotypes that were identical between the two data sets.

Comparing Population Genetic Summary Statistics between Genotyping Arrays

To compare the performances of the validated approximately 130,000 autosomal and X-chromosomal SNPs of the GenoChip array to commercial arrays, we obtained the list of SNPs for the Illumina Human660W-Quad BeadChip (544,366 SNPs) from Illumina and the Affymetrix Axiom Human Origins array (627,719 SNPs) available at ftp://ftp.cephb.fr/hgdp_supp10/Harvard_HGDP-CEPH/all_snp.map.gz (last accessed May 19, 2013). Because of the lack of overlap between these genotyping arrays, we used subsets of data calculated for HapMap III populations. Minor allele frequency (MAF) and F_{ST} estimates for African, European, and Asians were obtained from the "continental" HapMap data set, as described in Elhaik (2012). Briefly, genotype data of 602 unrelated individuals from eight populations (YRI, LWK, Maasai in Kinyawa, Kenya [MKK], CEU, TSI, CHB, Chinese from metropolitan Denver, Colorado [CHD], and JPT) were downloaded from the International HapMap Project web site (phase 3, second draft) (Altshuler et al. 2010), passed through rigorous filtering criteria, and finally merged into continental populations (African [288], European [144], and Asian [170]). The final continental data set consisted of 3 million SNPs genotyped in at least one population from each continent.

The MAF and F_{ST} values of the continental data set for autosomal (2,823,367) and X-chromosomal (86,449) SNPs were compared with those obtained from GenoChip (126,425 and 2,421 SNPs, respectively), Illumina Human660W (541,104 and 12,916 SNPs, respectively), and Affymetrix Axiom Human Origins Array (308,949 and 2,984 SNPs, respectively).

Because of the large number of F_{ST} values in each data set, their length distributions are very noisy. We thus adopted a simple smoothing approach in which F_{ST} values are sorted and divided to 1,000 equally sized subsets. The distribution of the mean F_{ST} value is then calculated using a histogram with 40 equally sized bins ranging from 0 to 1. To test whether two such F_{ST} distributions obtained by different arrays are different, we used the Kolmogorov–Smirnov goodness-of-fit test and the false discovery rate correction for multiple tests (Benjamini and Hochberg 1995). Because the differences between the distributions were highly significant due to the large sample sizes, we also calculated the effect size, first by using the nonoverlapping percentage of the two distributions, and then by using Hedges' g estimator of Cohen's d (Hedges 1981). If the area overlap is larger than 98% and Cohen's d is smaller than 0.05, we consider the magnitude of the difference between the two distributions to be too small to be biologically meaningful.

Principal components analysis (PCA) calculations were carried out using smartpca of the EIGENSOFT package (Patterson et al. 2006). Polygons were drawn manually around populations clustered separately from one another.

Results and Discussion

Designing the GenoChip

Choosing the Markers

The GenoChip was designed as an Illumina iSelect HD custom genotyping bead array that offers the ability to interrogate almost any SNP. In designing the chip, we endeavored to identify the fewest possible SNPs that offer an increased power for ancestry inference in comparison to random markers (Royal et al. 2010). SNPs that discern and identify populations are termed AIMs and are considered invaluable tools in population genetics and genetic anthropology. Half of our AIMs were culled from the literature, and the remaining were calculated using our novel AIMsFinder based on an approach described by Elhaik (2013) and infocalc (Rosenberg 2005) (supplementary text S1, Supplementary Material online). These two methods were applied on global panels comprising over 300 populations (supplementary table S1, Supplementary Material online) assembled from public and private data sets that were genotyped on a diversified set of arrays ranging from 30,000 to more than million SNPs in size. Many of these populations are unique to our project and have never before studied or searched for AIMs. Because AIMsFinder infers the minimal number of markers necessary to discern two genetically distinct populations, it was applied in a pairwise fashion over all the population data sets. In contrast, infocalc that ranks SNPs by their informativeness to ancestry was applied to whole population panels organized by the source of the genotype data (supplementary table S1, Supplementary Material online), where the top 1% of the results was considered AIMs. Overall, we ascertained over 80,000 autosomal and X-chromosomal AIMs from over 450 worldwide populations (fig. 1).

To facilitate studies on the extent of gene flow from Neanderthal and Denisovan to modern humans, we collected from the literature SNPs and haplotypes from genomic regions bearing evidence of interbreeding (Noonan et al. 2006; Green et al. 2010; Yotova et al. 2011). In addition, we used a

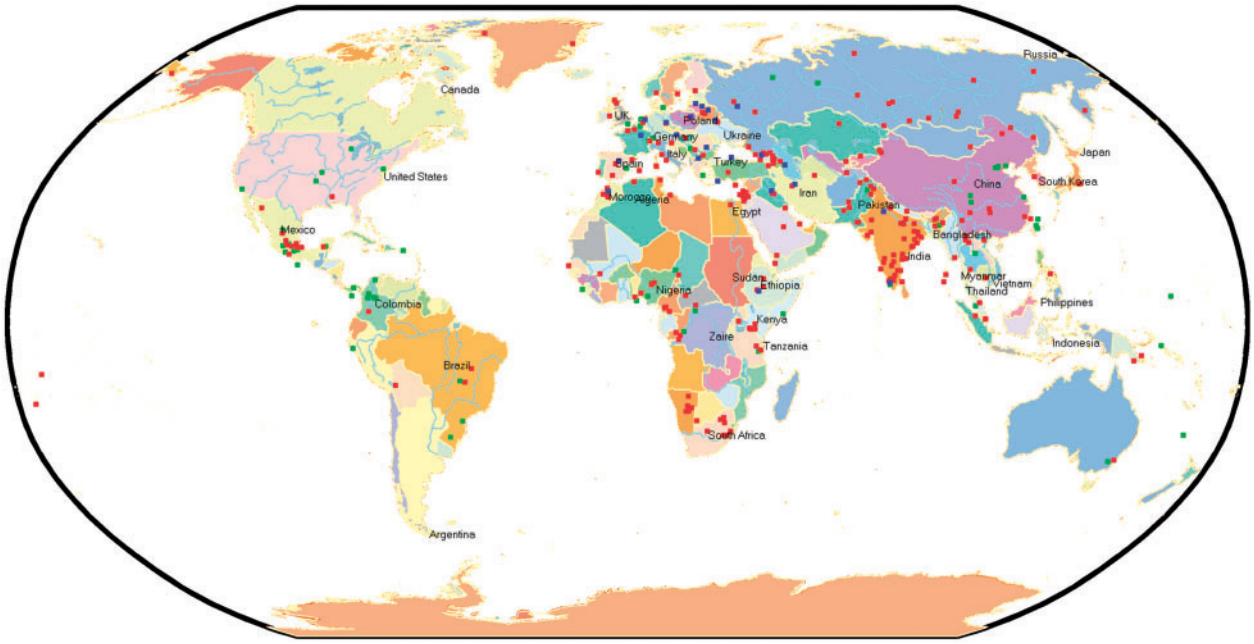

Fig. 1.—Worldwide distribution of population from which AIMs were obtained. AIMs from over 450 world populations were harvested from the literature (green) and calculated based on genotyped data from public and private collections (red) including over 30 Jewish populations (blue).

modified version of IsoPlotter+ (Elhaik et al. 2010; Elhaik and Graur 2013) to identify regions in which modern humans and Neanderthals share the derived allele and chimpanzees and Denisovans share the ancestral allele (supplementary text S1, Supplementary Material online). Using the same approach, we identified SNPs within regions enriched for the Denisovan shared derived alleles with humans. Overall, we included nearly 26,000 autosomal and X-chromosomal SNPs from potential interbreeding hotspots with extinct hominins. To support studies of more recent gene flow from ancient to modern humans, we included approximately 10,400 high-confidence Paleo-Eskimo Saqqaq SNPs (Rasmussen et al. 2010). In addition, we included approximately 12,000 high-confidence Aboriginal SNPs (Rasmussen et al. 2011). High-linkage disequilibrium (LD) SNPs ($r^2 > 0.4$) were excluded in all populations, by choosing a random SNP of the high-LD pair, except for hunter gatherers such as the Hadza and Sandawe of Tanzania (Tishkoff and Williams 2002) and Melanesian populations (Conrad et al. 2006) that are used to infer interbreeding with extinct hominins (Reich et al. 2010; Lachance et al. 2012).

To support potential imputation efforts, we supplemented regions of low SNP density (<1 SNP over 100,000 bases) with random common SNPs from HapMap III (1,000 SNPs with MAF > 20%) and the 1000 Genomes Project (3,500 SNPs with MAF > 10% in at least one continental population). To prevent false positives, we included mostly SNPs observed in both the HapMap III and 1000 Genome Project data sets (Altshuler et al. 2010; Durbin et al. 2010). We further eliminated A/T and C/G SNPs to minimize strand misidentification.

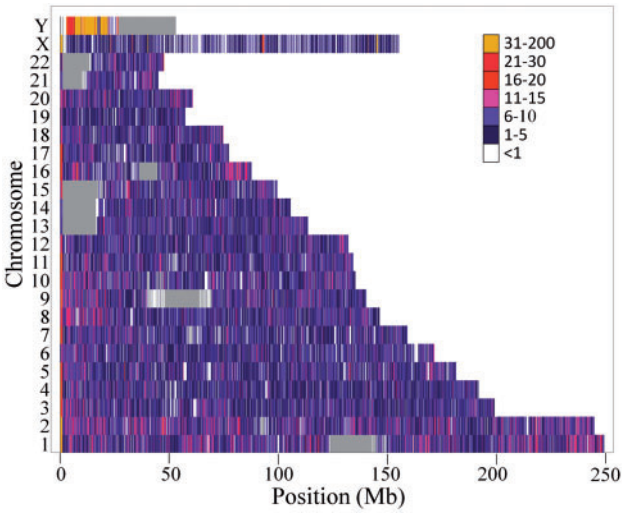

Fig. 2.—SNP density in the Genochip. The average numbers of GenoChip SNPs per 100,000 nucleotides across the genome are color coded. Gaps in the assembly are shown in gray.

The resulting chip has a SNP density of at least 1/100 kilobases over 92% of the assembled human genome (hg19) (fig. 2), including regions uncharted by the HapMap (I-III) and HGDP projects (Conrad et al. 2006; Altshuler et al. 2010). This high density of the chip and the excess inclusion of AIMs make it suitable for imputation, particularly for common markers (Pasaniuc et al. 2012).

Finally, we constructed over 45,000 probes to identify SNPs defining all known Y-chromosome and mtDNA haplogroups, many of which were not reported in the literature (supplementary text S2, Supplementary Material online).

Compatibility to Commercial Genotyping Arrays

Looking at autosomal and X-chromosomal SNPs, the GenoChip is highly compatible with other commercial arrays. Some 76% of our SNPs overlap with those in the Illumina Human 660W-Quad array, 55% overlap with the Illumina HumanOmni1-Quad, Illumina Express, and Affymetrix 6.0 arrays, and 40% overlap with the Affymetrix 5.0 and Affymetrix Human Origins arrays. With the exception of dedicated Y chromosome and mtDNA chips, the GenoChip includes the most comprehensive collection of uniparental markers.

Vetting the Chip for Health, Trait, or Medical Markers

Several steps were taken to ensure that the genetic results would not be exploited for pharmaceutical, medical, and biotechnological purposes. First, participant samples were maintained in complete anonymity during GenoChip analysis. Second, no phenotypic or medical data were collected from the participants. Third, we included only SNPs in noncoding regions without any known functional association (Graur et al. 2013), as reported in dbSNP build 132. Last, we filtered our SNP collection against a 1.5 million SNP data set (Pheno SNPs) containing all variants that have potential, known, or suspected associations with diseases.

To construct the Pheno SNPs data set, we extracted SNPs from multiple open-access databases including the Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim/, last accessed May 19, 2013), the Cancer Genome Atlas (Hudson et al. 2010), PhenCode (Giardine et al. 2007), the National Human Genome Research Institute (NHGRI) Genome-Wide Association Studies (GWAS) Catalog (Hindorff et al. 2009), The Genetic Association Database (Becker et al. 2004), MutaGeneSys (Stoyanovich and Pe'er 2008), GWAS Central (Thorisson et al. 2009), and SNPedia, as well as SNPs identified in the major histocompatibility complex (MHC) region. We also excluded SNPs reported to be associated with phenotypic traits. Finally, to circumvent imputation efforts toward inferring potential medical-relevant SNPs, we excluded SNPs that were in high LD ($r^2 > 0.8$) with the Pheno SNPs.

We thus designed the first genotyping array dedicated for genetic anthropological and genealogical research that is suitable for detecting gene flow from archaic hominins and ancient humans into modern humans as well as between worldwide populations. The final GenoChip has over 130,000 highly informative autosomal and X-chromosomal markers, approximately 12,000 Y-chromosomal markers, and approximately 3,300 mtDNA markers without any known health, medical, or phenotypic relevance (supplementary table S2, Supplementary Material online).

Validating the GenoChip Results

The accuracy of the autosomal genotypes obtained by the GenoChip was assessed by genotyping 168 worldwide samples from the 1000 Genomes Project and cross-validating the results. The concordance rate per sample was over 99.5%. We did not observe any position with mismatching homozygote alleles. The marginal error rate was expected due to the low coverage of the 1000 Genomes Project data, particularly for rare alleles (Durbin et al. 2010). We thus confirmed that genotypes reported by the GenoChip are accurate.

The ability of the GenoChip to infer uniparental haplogroups was similarly assessed by genotyping 400 additional samples with known haplogroups. The haplotypes of these samples were confirmed by Sanger sequencing of the full mitochondrial genome and all relevant Y chromosome SNP locations that determined the exact haplogroup down to the last branch of the published Y-chromosomal tree (supplementary text S2, Supplementary Material online). The average success rates for the paternal and maternal haplogroups were 82% and 90%, respectively (fig. 3). The reasons for our inability to validate the remaining haplogroups are the unavailability of control samples to identify deeper splits in the tree. Moreover, some haplogroups cannot be measured with the Illumina bead chip technology because they are not represented by a real SNP but rather by large-scale variations of repetitive elements. We note that some of the failed markers for particular haplogroups can be substituted by phylogenetically equivalent markers and rescue these haplogroups, although formally they were counted as missing. Our experience with the tens of thousands of GenoChip participants indicates that most samples (>99%) are classified on haplogroup branches that are perfectly captured by the GenoChip. The remaining users for which the exact position along the tree cannot be assigned (e.g., R-P312*) are classified to a higher level haplogroup (e.g., R-P310). A large-scale genotyping effort to validate the remaining haplogroups is undergoing. We thus confirmed that GenoChip produces highly accurate results and has broad coverage for markers defining Y-chromosome and mtDNA haplogroups.

Testing the GenoChip's Abilities to Discern Populations

MAF Distribution

Before comparing the ability of the GenoChip SNPs to discern populations, we compared the similarity of their MAF distribution with those of the Illumina Human660W and Affymetrix Human Origins SNP arrays. Because of the low overlap of these three arrays, we obtained and analyzed genotype data from eight HapMap populations. The results of the complete set of HapMap markers were compared with three subsets of markers that overlapped with those of each array.

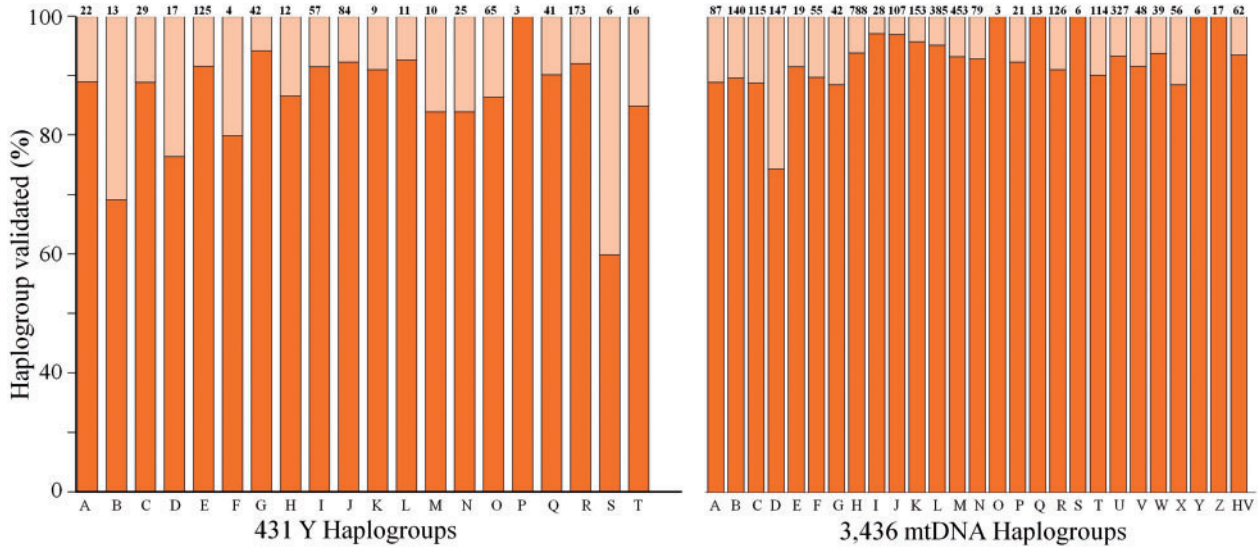

Fig. 3.—Success rate in identifying Y-chromosomal (left) and mtDNA (right) haplogroups. The plots depict all known basal haplogroups (columns), the number of known subgroups in each haplogroup (top of each column), and the proportion of subgroups that were validated with the GenoChip.

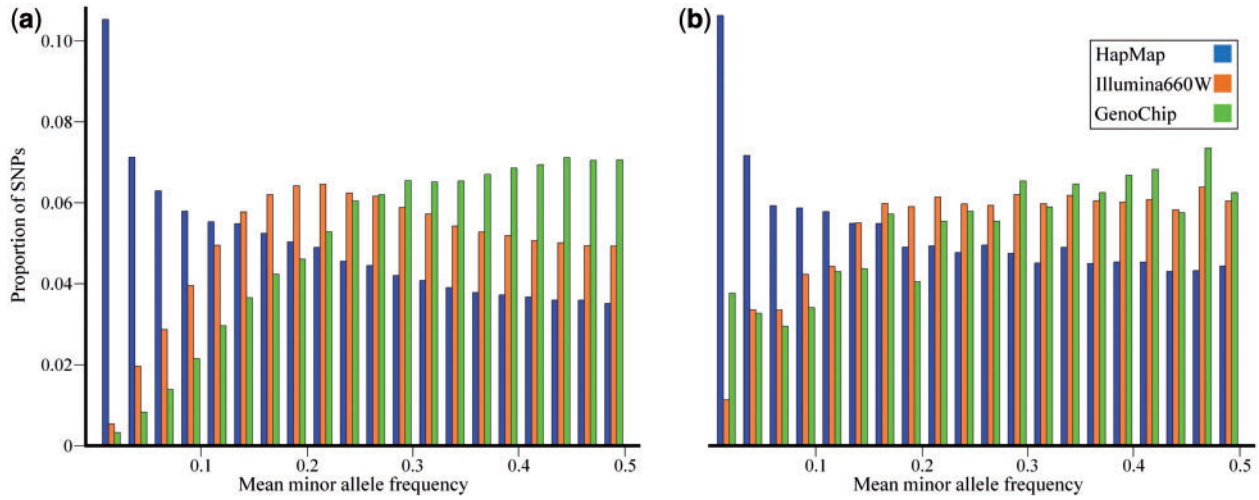

Fig. 4.—MAF distributions for autosomal (*a*) and X-chromosomal (*b*) HapMap SNPs. MAF distributions are shown for HapMap SNPs and two subsets that overlap with the Illumina Human660W and GenoChip SNPs.

A comparison of the MAF distributions of the three arrays revealed gross differences in allele frequencies (fig. 4, supplementary fig. S1, Supplementary Material online). In the HapMap data set, over 82% of the SNPs are common (MAF > 0.05) and less than 5% are considered rare (MAF < 0.01). The proportion of common SNPs in all the arrays is similar (96–98%), but the GenoChip is enriched for the most common SNPs (MAF > 0.25). Because of the high frequency of the rare ENCODE SNPs in the HapMap data set, none of the arrays resembled the shape of the HapMap's MAF distribution. Nonetheless, both the Human660W (0.07%) and Human Origins (0.36%) arrays are enriched in rare SNPs compared with the GenoChip (0.008%). Similar trends were observed for X-chromosomal SNPs. Here, the HapMap data set consisted of 83% common SNPs, compared with 93% for the GenoChip and 96% for the commercial arrays. The GenoChip array exhibits similar enrichment in the most common SNPs (MAF > 0.3), but unlike the commercial arrays, it also consists of 1% extremely rare SNPs due to the inclusion of rare haplotypes speculated to indicate interbreeding with archaic hominins. Altogether, the MAF distributions of the three arrays differ from the HapMap MAF distribution and

correspond to the choices of SNP ascertainment made in the design of each array.

Genomewide F_{ST} Distribution

To assess the extent of genetic diversity that can be inferred among human subpopulations by the different arrays, we next compared their F_{ST} distributions (Wright 1951). F_{ST} measures the differentiation of a subpopulation relative to the total population and is directly related to the variance in allele frequency between subpopulations, such that a high F_{ST} corresponds to a larger difference between subpopulations (Holsinger and Weir 2009). Elhaik (2012) used 1 million markers that were genotyped in 602 HapMap samples from eight populations to carry out a two-level hierarchical F_{ST} analysis. He showed that the greatest proportion of genetic variation occurred within individuals residing in the same populations, with only a small amount (12%) of the total genetic variation being distributed between continental populations and even a lesser amount (1%) between intracontinental populations. An F_{ST} distribution for three continental populations employing 3 million HapMap SNPs yielded an even lower estimate (8%) to the proportion of genetic variation distributed between continental populations due to the large number of rare alleles (Elhaik 2012).

In a similar manner to (Elhaik 2012) later analysis, we used the F_{ST} values calculated for eight HapMap populations grouped into three continental populations to create three subsets for the markers that overlap with each array. Although all F_{ST} distributions were similar in shape to the HapMap F_{ST} distribution, they differed in their means (fig. 5, supplementary fig. S2, Supplementary Material online). The autosomes and X-chromosomal SNPs of the commercial arrays have significantly lower F_{ST} values (Kolmogorov–Smirnov goodness-of-fit test, $P < 0.05$) than that of the GenoChip due to the high fraction of rare uninformative SNPs in these arrays. The magnitude of the differences between the F_{ST} values of the GenoChip to those of the commercial arrays were also large for autosomal (area overlap 86–91%, Cohen's *d* 0.09–0.13) and X-chromosomal SNPs (area overlap 93%, Cohen's *d* 0.09–0.11). These results suggest a reduced ability of the commercial arrays to elucidate ancient demographic processes (Kimura and Ota 1973; Watterson and Guess 1977).

The Illumina Human660W array had the highest fraction of low-F_{ST} alleles, suggesting it is the least suitable for population genetic studies compared with the GenoChip and Human Origins. As only half of the Human Origins SNPs could be tested, it is difficult to evaluate its performance. However, we speculate that the large number of rare alleles reflect the private alleles of the dozen populations used for its ascertainment. Because the MAF and F_{ST} were not used as filtering criteria for the GenoChip SNPs, we can conclude that its enrichment toward high-F_{ST} SNPs mirrors the success of the ascertainment process and its potential for population genetic studies.

Genetic Diversity in Worldwide Populations

Last, PCA (Price et al. 2006) was used to explore the extent of population differentiation between 14 worldwide populations that were genotyped on the GenoChip in the validation stage (fig. 6*A*). The samples aligned along the two well-established geographic axes of global genetic variation: PC1 (sub-Saharan Africa vs. the rest of the Old World) and PC2 (east vs. west Eurasia) (e.g., Li et al. 2008; Elhaik 2013). GenoChip results reveal geographically refined groupings of Eastern (Luhya) and Western (Yoruba) Africans, Eastern (Chinese and Japanese) and South Eastern (Vietnamese) Asians, Amerindian (Peruvians Mexicans) and Indian populations, and finally

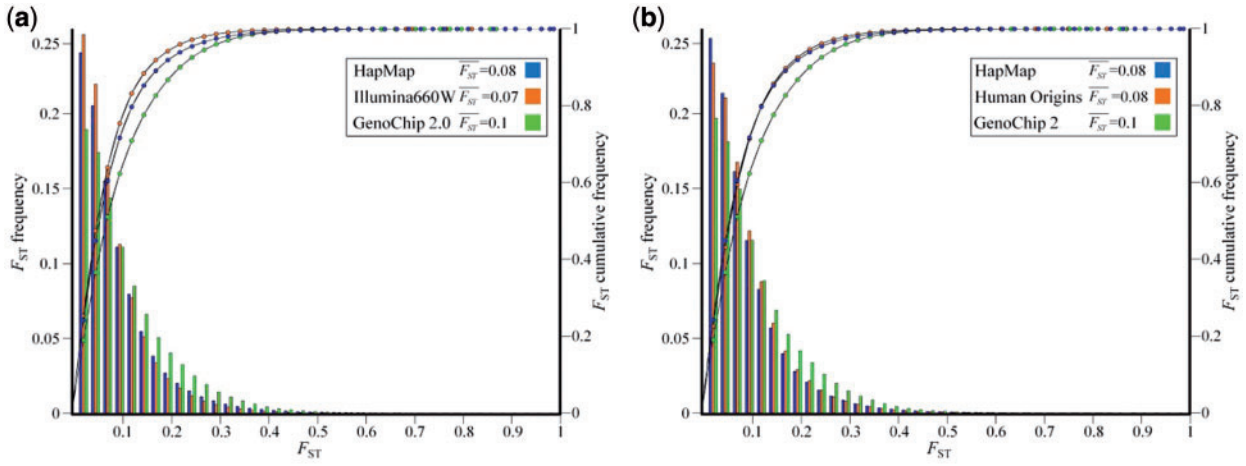

Fig. 5.—Distribution of locus-specific F_{ST} in three continental populations. F_{ST} values were obtained for (*a*) autosomal and (*b*) X-chromosomal HapMap SNPs. F_{ST} distributions are shown for HapMap SNPs and two subsets that overlap with the Illumina Human660W and GenoChip SNPs. The histograms show bin distribution as indicated on the *x* axis and the cumulative distribution (line).

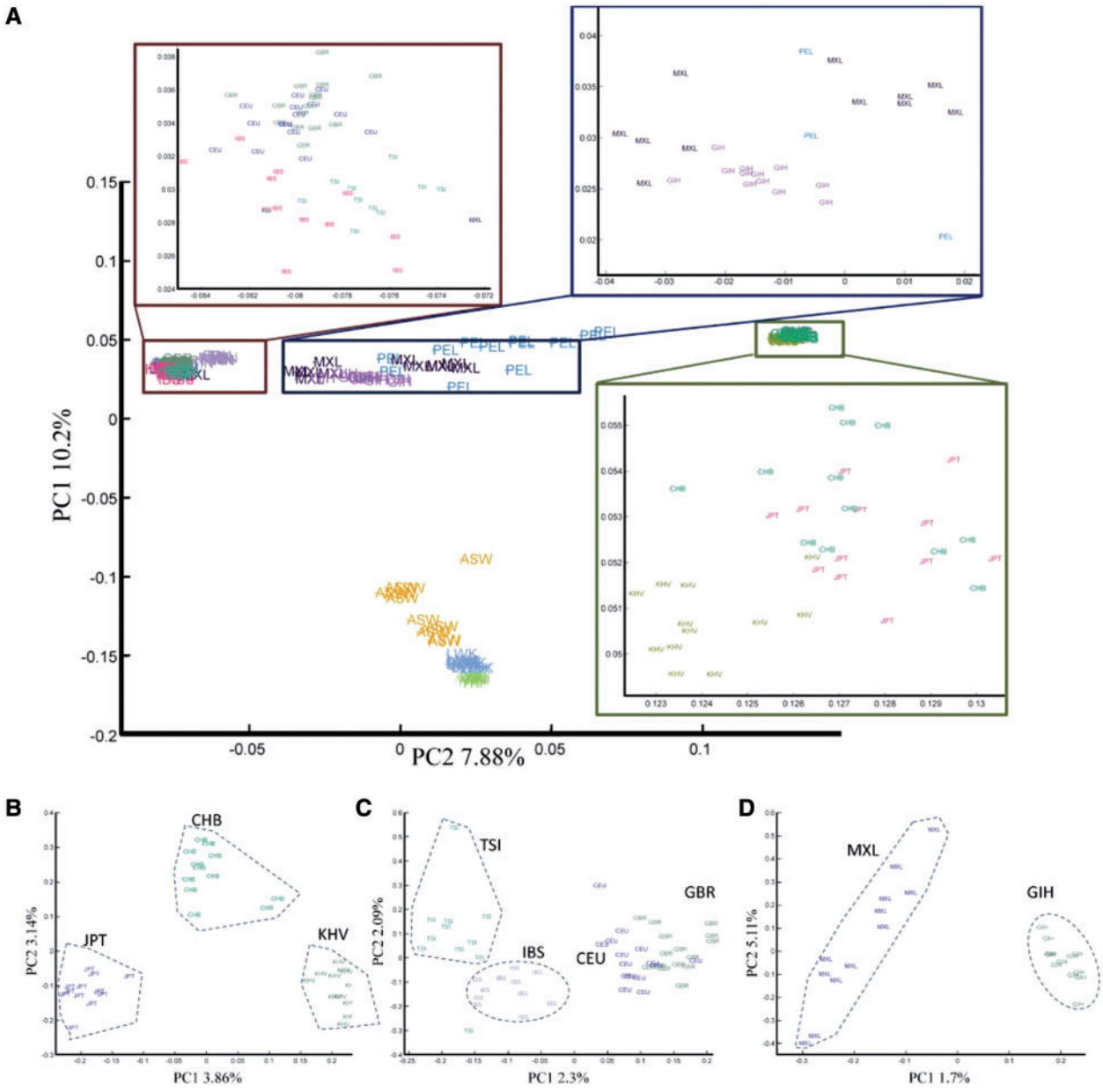

Fig. 6.—PCA plots of genetic diversity across 14 worldwide populations. Each figure represents the genetic diversity seen across the populations considered, with each sample mapped onto a spectrum of genetic variation represented by two axes of variations corresponding to two eigenvectors of the PCA. Individuals from each population are represented by a unique color. (*A*) Analysis of all populations. The insets magnify European, Asian, and the cluster of Amerindian and Indian individuals. (*B*) Analysis of East Asian individuals. (*C*) Analysis of European individuals. (*D*) Analysis of Amerindian and Indian individuals. A polygon surrounding all or most of the individuals belonging to a group designation highlights the population groups.

Northern (Finnish), Southern (Italian and Iberians), and Western (British and CEU) Europeans. As expected, the Amerindian populations form a gradient along the diagonal line between European and East Asians based on their dominant ancestry as did the African Americans along the diagonal line between Africans and Europeans. These patterns are similar to those observed in worldwide populations using commercial arrays (e.g., Teo et al. 2009; Xing et al. 2010).

When we consider only the East Asian populations (comprising CHB, JPT, and KHV), the first and second axes of variation completely separated the three populations (fig. 6*B*), in agreement with Teo et al. (2009). In a similar manner, we were able to differentiate Gujarati Indians and Americans of Mexican ancestry (fig. 6*C*), as well as Italians, Iberians, and Western European populations (fig. 6*D*), with the exception of one TSI outlier. As expected, some overlap

was observed between individuals of Northern and Western European ancestry (CEU) and British (GBR).

Conclusions

To summarize, we designed, developed, validated, and tested the GenoChip, the first genotyping chip completely dedicated to genetic anthropology. The GenoChip will help to clarify the genetic relationships between archaic hominins such as Neanderthal and Denisovan, extinct humans, and modern humans as well as to provide a more detailed understanding of human migratory history. We compared the MAF and F_{ST} distributions of the GenoChip SNPs to those of HapMap and two commercially available arrays and demonstrated the ability of the GenoChip to differentiate subpopulations within global data sets. We expect that the expanded use of the GenoChip in genetic anthropology research will expand our knowledge of the history of our species.

Supplementary Material

Supplementary text S1 and S2, tables S1 and S2, and figures S1–S4, and are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to David Reich, Nick Patterson, Morten Rasmussen, Robert Hastings, and Dienekes Pontikos for sharing their data with us and for fruitful discussions. They also thank Alon Keinan and the Illumina development team for their feedback and support. This work was supported by the National Geographic Society, by SAR-LR 7/09, cRP2-597 to P.F., and by The Wellcome Trust (098051) to C.T.S. and Y.X.

The Genographic Consortium (members are listed alphabetically by surname): Oscar Acosta (Universidad San Martin de Porres, Lima, Peru), Syama Adhikarla (Madurai Kamaraj University, Madurai, Tamil Nadu), Christina J. Adler (University of Adelaide, South Australia, Australia), Elena Balanovska (Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow), Oleg Balanovsky (Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow), Jaume Bertranpetit (Universitat Pompeu Fabra, Barcelona, Spain), Andrew C. Clarke (University of Otago, Dunedin, New Zealand), David Comas (Universitat Pompeu Fabra, Barcelona, Spain), Alan Cooper (University of Adelaide, South Australia, Australia), Matthew C. Dulik (University of Pennsylvania, Philadelphia, Pennsylvania), Jill B. Gaieski (University of Pennsylvania, Philadelphia, Pennsylvania), Arun Kumar Ganesh Prasad (Madurai Kamaraj University, Madurai, Tamil Nadu), Wolfgang Haak (University of Adelaide, South Australia, Australia), Marc Haber (Lebanese American University, Chouran, Beirut), Matthew E. Kaplan (University of Arizona, Tucson, Arizona), Daniela R. Lacerda (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais), Shilin Li (Fudan University, Shanghai, China), Begoña Martínez-Cruz (Universitat Pompeu Fabra, Barcelona, Spain), Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand), Nirav C. Merchant (University of Arizona, Tucson, Arizona), John R. Mitchell (University of Pennsylvania, Philadelphia, Pennsylvania), Amanda C. Owings (University of Pennsylvania, Philadelphia, Pennsylvania), Laxmi Parida (IBM, Yorktown Heights, New York), Ramasamy Pitchappan (Madurai Kamaraj University, Madurai, Tamil Nadu), Daniel E. Platt (IBM, Yorktown Heights, New York), Lluis Quintana-Murci (Institut Pasteur, Paris, France), Colin Renfrew (University of Cambridge, Cambridge, United Kingdom), Ajay K. Royyuru (IBM, Yorktown Heights, New York), Jose Raul Sandoval (Universidad San Martin de Porres, Lima, Peru; Universidade Federal de Minas Gerais, Belo Horizonte, Brazil), Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu), Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais), Clio S. I. Der Sarkissian (University of Adelaide, South Australia, Australia), Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa), David F. Soria Hernanz (National Geographic Society, Washington, District of Columbia), Pandikumar Swamikrishnan (IBM, Somers, New York), Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), Miguel G. Vilar (University of Pennsylvania, Philadelphia, Pennsylvania), Pierre A. Zalloua (Lebanese American University, Chouran, Beirut), Janet S. Ziegle (Vitapath Genetics, Foster City, California).

Literature Cited

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 27: 2534–2547.

Altshuler DM, et al. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.

Balanovsky O, et al. 2011. Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol. 28:2905–2920.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet. 36:431–432.

Behar DM, et al. 2010. The genome-wide structure of the Jewish people. Nature 466:238–242.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B. 57:289–300.

Burger J, Strong MF. 1990. The Gaia atlas of first peoples: a future for the indigenous world. New York: Doubleday.

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cavalli-Sforza LL. 2007. Human evolution and its relevance for genetic epidemiology. Annu Rev Genomics Hum Genet. 8:1–15.

Chaubey G, et al. 2011. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol. 28:1013–1024.

Conrad DF, et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 38: 1251–1260.
Durbin RM, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Elhaik E. 2012. Empirical distributions of F_{ST} from large-scale human polymorphism data. PLoS One 7:e49837.
Elhaik E. 2013. The missing link of Jewish European ancestry: Contrasting the Rhineland and the Khazarian hypotheses. Genome Biol Evol. 5: 61–74.
Elhaik E, Graur D. 2013. IsoPlotter+: A tool for studying the compositional architecture of genomes. ISRN Bioinformatics 2013:6.
Elhaik E, Graur D, Josic K, Landan G. 2010. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acids Res. 38:e158.
Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci U S A. 109: 13956–13960.
Fardon R. 2012. The Sage handbook of social anthropology. Thousand Oaks (CA): SAGE Publications.
Florez JC, et al. 2009. Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. Diabetologia 52:1528–1536.
Giardine B, et al. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat. 28:554–562.
Graur D, et al. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 5:578–590.
Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.
Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat. 29:648–658.
Hatin WI, et al. 2011. Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. PLoS One 6:e18312.
Hedges LV. 1981. Distribution theory for glass's estimator of effect size and related estimators. J Educ Behav Stat. 6:107–128.
Henn BM, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 108:5154–5162.
Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. 2010. Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet. 19:R221–R226.
Hindorff LA, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 106:9362–9367.
Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet. 10:639–650.
Hudson TJ, et al. 2010. International network of cancer genome projects. Nature 464:993–998.
Hunter-Zinck H, et al. 2010. Population genetic structure of the people of Qatar. Am J Hum Genet. 87:17–25.
Kidd JR, et al. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet. 2:1.
Kimura M, Ota T. 1973. The age of a neutral mutant persisting in a finite population. Genetics 75:199–212.
Klein RG. 2008. Out of Africa and the evolution of human behavior. Evol Anthropol Issues News Rev. 17:267–281.
Kosoy R, et al. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 30:69–78.
Lachance J, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150:457–469.
Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5: e10284.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.
Lu Y, Patterson N, Zhan Y, Mallick S, Reich D. 2011. Technical design document for a SNP array that is optimized for population genetics. ftp://ftp.cephb.fr/hgdp_supp10/8_12_2011_Technical_Array_Design_Document.pdf (last accessed May 19, 2013).
McEvoy BP, et al. 2009. Geographical structure and differential natural selection among North European populations. Genome Res. 19: 804–814.
McEvoy BP, et al. 2010. European and Polynesian admixture in the Norfolk Island population. Heredity 105:229–234.
Nassir R, et al. 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet. 10:39.
Noonan JP, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113–1118.
Pasaniuc B, et al. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 44:631–635.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.
Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. PLoS Genet. 2:e105.
Price AL, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.
Price AL, et al. 2007. A genomewide admixture map for Latino populations. Am J Hum Genet. 80:1024–1036.
Price AL, et al. 2008. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 4:e236.
Rasmussen M, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762.
Rasmussen M, et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334:94–98.
Reich D, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.
Reich D, et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 89: 516–528.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.
Rosenberg NA. 2005. Algorithms for selecting informative marker panels for population assignment. J Comput Biol. 12:1183–1201.
Royal CD, et al. 2010. Inferring genetic ancestry: opportunities, challenges, and implications. Am J Hum Genet. 86:661–673.
Schurr TG, et al. 2012. Clan, language, and migration history has shaped genetic diversity in Haida and Tlingit populations from Southeast Alaska. Am J Phys Anthropol. 148:422–435.
Silva-Zolezzi I, et al. 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proc Natl Acad Sci U S A. 106:8611–8616.
Stoyanovich J, Pe'er I. 2008. MutaGeneSys: estimating individual disease susceptibility based on genome-wide SNP array data. Bioinformatics 24:440–442.

Stringer CB, Andrews P. 1988. Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268.
Teo YY, et al. 2009. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. 19:2154–2162.
Thorisson GA, et al. 2009. HGVbaseG2P: a central genetic association database. Nucleic Acids Res. 37:D797–D802.
Tian C, et al. 2008. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet. 4:e4.
Tian C, et al. 2009. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. Mol Med. 15:371–383.
Tishkoff SA, Williams SM. 2002. Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet. 3:611–621.
Watterson GA, Guess HA. 1977. Is the most frequent allele the oldest? Theor Popul Biol. 11:141–160.
Wolinsky H. 2006. Genetic genealogy goes global. Although useful in investigating ancestry, the application of genetics to traditional genealogy could be abused. EMBO Rep. 7:1072–1074.
Wright S. 1951. The genetical structure of populations. Ann Eugen. 15: 323–354.
Xing J, et al. 2009. Fine-scaled human genetic structure revealed by SNP microarrays. Genome Res. 19:815–825.
Xing J, et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96:199–210.
Yang N, et al. 2005. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet. 118:382–392.
Yotova V, et al. 2011. An x-linked haplotype of Neandertal origin is present among all non-African populations. Mol Biol Evol. 28: 1957–1962.
Yunusbayev B, et al. 2012. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol. 29:359–365.

Associate editor: Dan Graur

Genome Biology and Evolution SMBE